A Split-Inference Intracortical Interface IC for Battery-Free mm-Scale Magnetoelectrically Powered Brain Implants

Mustafa Kanchwala*[1], Jianxiong Xu*[1], Mohammad Abdolrazzaghi*[1], Wonjune Kim[2], Gerard O'Leary[1], Yu Huang[1], Junyu Ma[1], Jose Sales Filho[1], Sudip Nag[1], Hanfeng Cai[1], Qiaosong Deng[1], Weian Deng[1], Anush Mutyala[1], Theeban Kumaresan[1], Shidu Ren[1], Chae Lim[1], Mandana Movahed[1], Homeira Moradi[1], George Eleftheriades[1], Taufik A. Valiante[1], Jacob T. Robinson[2], Roman Genov[1]

[1]University of Toronto, Toronto, ON, Canada; [2]Rice University, Houston, TX, USA, *Equal contribution authors

Abstract

This paper presents a brain-interfacing IC that enables depth-scalable edge inference within the tight form factor and energy budget of a minimally invasive, battery-free implant. A 181-dB FoM noise-shaping SAR ADC with an LNA is shared across 20 of 32 recording electrodes. On-chip inference combines boosted decision-stump spike classification with first-layer spiking- and deep-neural-network (SNN/DNN) inference, reducing the output data rate by up to 16,732×. An event-based body-coupled communication (BCC) TX streams intermediate activations to a wearable device for remaining-layer execution. The communication and computation energy costs are 1.23 nW and 53 nW per input, respectively, enabling wireless powering by a mm-scale magnetoelectric film.

Introduction

Smart brain–interfacing devices, such as closed-loop neurostimulators and brain–machine interfaces (BMIs), link biological neural networks with machine-learning systems and hold promise for improving outcomes in neurological disorders, including movement and speech deficits following stroke or spinal cord injury. However, their intelligence is strongly constrained by strict heat-dissipation limits: on-device inference is energetically expensive [1], while remote inference incurs high wireless communication energy costs, necessitating data-reduction schemes whose additional computation can offset the benefits of offloading [2-4].

Minimal invasiveness further necessitates battery-free wireless power transfer. Low-frequency (LF) inductive powering penetrates tissue efficiently but requires large receiver coils [5], whereas high-frequency inductive powering enables mm-scale implants at the cost of higher tissue absorption and sensitivity to misalignment [6]. Ultrasound avoids these trade-offs but can be attenuated by the skull [7], motivating magnetoelectric (ME) powering [8,9], which converts deeply penetrating LF magnetic fields into electrical energy via magnetostrictive–piezoelectric transduction.

System Architecture and Implementation

The intracortical interface IC, targeting minimally invasive periodic neuro-therapy applications, is shown in Fig. 1 (top). Inference is partitioned into local (on-implant) and remote (wearable) stages, with intermediate results transmitted at data rates reduced by up to 16,732× relative to raw data. This split-inference architecture enables deep end-to-end trained models via local partial inference, without incurring the high energy costs of exclusively local or remote inference. The local stage dissipates 53 nW/input by performing boosted decision-stump spike classification followed by first-layer inference using either a spiking neural network (SNN) or a deep neural network (DNN), targeting motor and inner-speech decoding, respectively. An event-based body-coupled communication (BCC) TX transmits sparse first-layer activations via time-division multiple access to a wearable FPGA shared by up to eight implants, where remaining layers are executed. The inference outputs trigger on-chip electrical stimulation via the wireless powering link for stroke rehabilitation or enable auditory feedback for guided inner-speech training.

The IC dissipates 6.07 μW/input during inference, enabling periodic powering by a chip-scale ME film, as shown in Fig. 1 (bottom). The target implantation regions – the motor cortex and Wernicke's speech area – lie along narrow bands of tissue, enabling power delivery from a custom alignment-controlled headphone headband. An array of selectable alternating-polarity coils (APCs) spatially focuses the magnetic field into deep tissue. A 3×1×0.3 mm³ ME film is stacked with a 1×2 mm² IC and a 1 mm² digital controller, and integrated into a flexible, mm-scale microelectrode shank with 16×2 dual-side electrodes. The battery-free shank floats with surrounding tissue and may reduce motion artifacts during powered therapy sessions compared with anchored electrodes.

Fig. 2 shows the AFE, comprising a neural LNA and a 2nd-order noise-shaping (NS) SAR ADC. The amplifier combines auto-zeroing and chopping to suppress offset-induced ripple and folded noise, and includes a DC servo loop to mitigate electrode drift, achieving >1 GΩ input impedance at DC. As shown in Fig. 3, the amplifier achieves 10.3 nVrms/√Hz input-referred noise (NEF = 1.5), while the entire AFE attains a competitive Schreier FoM of 181 dB and 15.8 ENOB over 8 kHz [6], [10-12]. For spike classification, 5.8 bits of ENOB and a 4× increase in power are traded for additional bandwidth, enabling multiplexing of 20 10-bit inputs from 32 electrodes.

Fig. 4 illustrates local inference within a dynamic power budget of 53 nW/input, without using multipliers. Building on evidence in [13] that exponential temporal filtering yields a task-agnostic, low-complexity neural representation, we use an exponentially weighted moving-average (EMA) filter to transform spike morphology into simple threshold-based temporal features efficiently exploited by shallow decision tree classifiers. Each AFE output channel is fed to two EMA filters (Fig. 4, top), to extract noise-robust, linearly separable spike features using only shift-and-add operations. These features are classified using a boosted decision-stump model trained off-chip with XGBoost, with subject- and task-specific parameters retrainable without hardware modification (EMA time scales α, comparator thresholds TH, and morphology-derived delays Δ), enabling classification of up to four neurons per input. Simplified first-layer SNN or DNN inference is executed locally, with remaining layers offloaded to the wearable device. Time-multiplexing of 20 AFE inputs enables classification of up to 80 biological neurons, as illustrated by the SNN example in Fig. 4 (right), where leaky integrate-and-fire (LIF) neurons activate based on input spike rate.

Communication energy reduction is demonstrated in two applications (Fig. 5). For motor decoding in non-human primates [14], a three-class split SNN classifies precision grip, side grip, and no grip with 90.4% cross-validated accuracy while reducing output data rate by 1,066×. For inner-speech decoding in humans [15], an eight-class split DNN identifies the most important neurons (56 of up to 80), achieving 96.1% per-class accuracy with a 16,732× data-rate reduction and supporting periodic retraining to mitigate foreign-body response. Figs. 6–8 show implementation and characterization of the BCC data TX and ME power management circuits. Figs. 9–10 and Tables I–III provide comparative analysis.

References

[1] M. A. Shaeri et al., ISSCC, 2024, pp. 546–548, doi: 10.1109/ISSCC49657.2024.10454533.

[2] Z. Wen et al., Symp. on VLSI, 2025, pp. 1–3, doi: 10.23919/VLSITechnologyandCir65189.2025.11074980.

[3] A. Akhoundi et al., ISSCC, 2025, pp. 268–270, doi: 10.1109/ISSCC49661.2025.10904780.

[4] Y. Chen et al., ISSCC, 2023, pp. 486–488, doi: 10.1109/ISSCC42615.2023.10067264

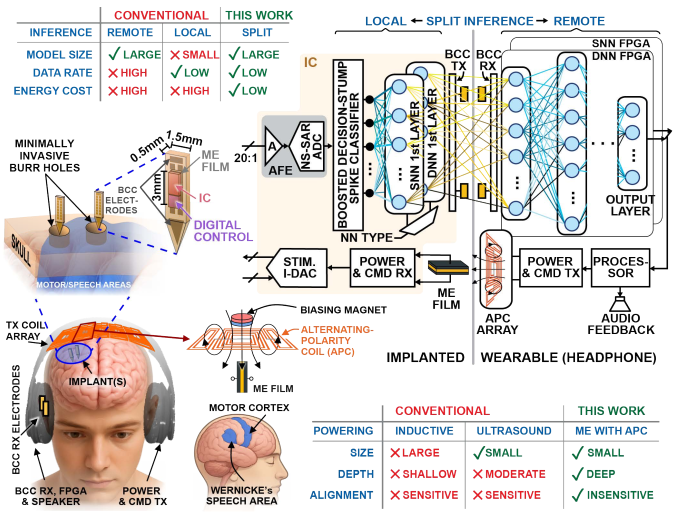


Fig. 1. Envisioned smart brain implant with inference split between the intracortical interface IC and a wearable headphone, using the magnetoelectric (ME) effect for powering and a body-coupled communication (BCC) link to transmit intermediate activations.

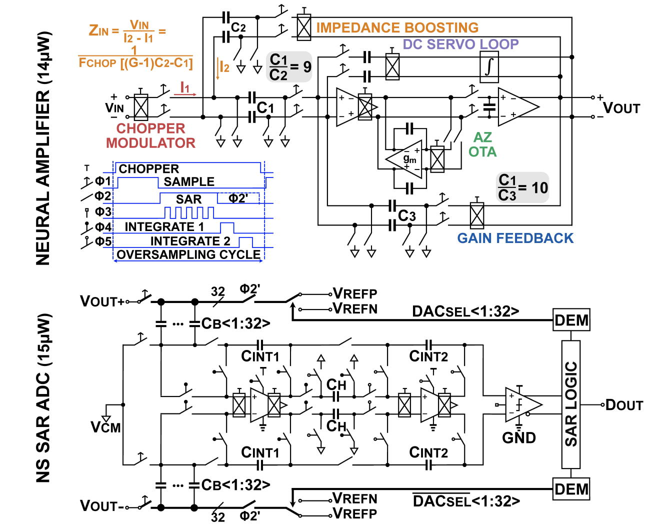


Fig. 2. Analog front-end (AFE) including a neural recording low-noise amplifier (LNA) (top) and a noise-shaping (NS) SAR ADC (bottom).

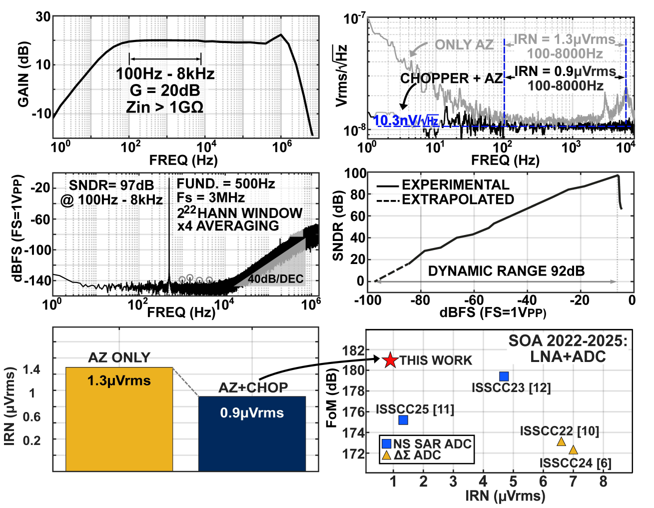


Fig. 3. Analog front-end experimentally measured results (top and middle), and comparison with state-of-the-art (SOA) neural recording analog front-ends, including both the LNA and the ADC (bottom).

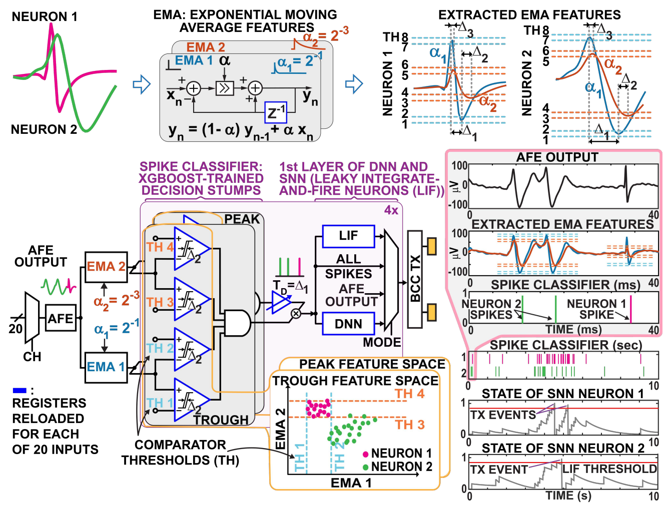


Fig. 4. Local on-chip inference digital back-end (DBE) comprising an EMA feature extractor, a decision-stump spike classifier, and SNN (LIF) or DNN first-layer neurons. SNN/DNN modes reuse the same local spike-classification and first-layer-inference hardware.

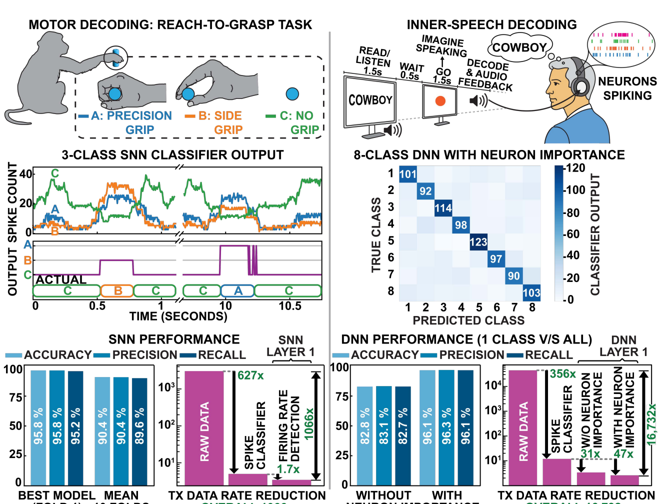


Fig. 5. Split-inference performance validation in motor (left) and inner-speech (right) decoding. Accuracy is reported on existing datasets [14,15] to illustrate achievable output data-rate reductions of 1,066× and 16,732× under representative neural decoding tasks.

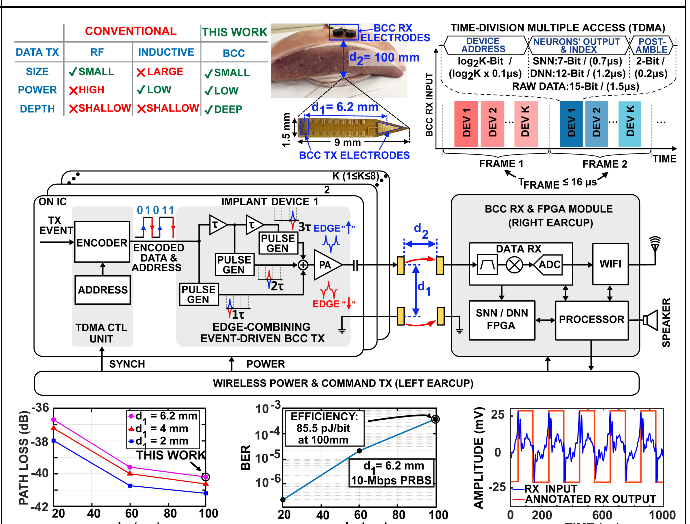
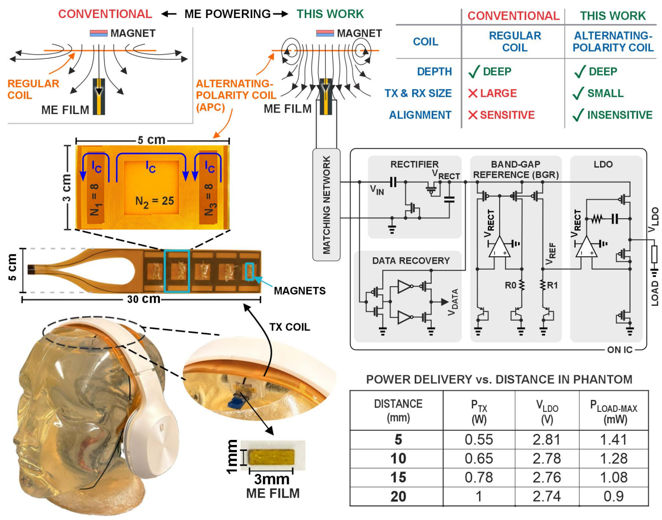


Fig. 6. Event-based body-coupled communication (BCC) data transmitter (TX) system diagram and test setup using ex vivo porcine tissue as a human tissue surrogate (top), and experimentally measured results obtained from the same setup (bottom).



POWER DELIVERY vs. DISTANCE IN PHANTOM

DISTANCE (mm)	P_{TX} (W)	V_{LDO} (V)	$P_{LOAD\text{-}MAX}$ (mW)
5	0.55	2.81	1.41
10	0.65	2.78	1.28
15	0.78	2.76	1.08
20	1	2.74	0.9

Fig. 7. Magnetoelectric (ME) wireless powering validation using a saline-filled head phantom (left), and on-chip power management circuits with experimentally measured results (right).

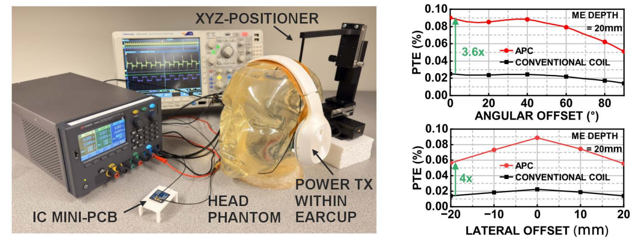


TABLE I. Wireless powering performance comparison.

	THIS WORK		JSSC24 [8]	ISSCC25 [9]	NAT. BME20 [7]	ISSCC21 [5]
TECHNOLOGY (nm)	65		180	180	65	180
WPT TYPE	ME WITH APC		ME W/ CONVEN. COIL		ULTRASOUND (US)	INDUCTIVE
FREQUENCY (kHz)	271		340	320	1850	6780
TX AREA (cm^2)	15		>50	>80	5	4.6
RX AREA (mm^2)	3		10	35	0.8	452
TX-RX DISTANCE (mm)	20		20	20	18	6.5
PEAK PTE (%)	0.09		0.55	0.3	0.06	61.9
MAX PDL (mW)	STIM. IDLE	STIM. ON	2.75	-	0.0097	32
	0.3	0.9				
FOM (mW/cm^3)*	**13.5**	**40**	11	-	4.4	9.97

* FOM=(PDL×DISTANCE)/(RX AREA×TX AREA) PDL: POWER DELIVERED TO LOAD PTE: POWER TRANSFER EFFICIENCY

Fig. 8. ME wireless powering test setup and experimental results on robustness to misalignment (top), and comparison with SOA wirelessly powered ICs (bottom). Power transfer is evaluated under controlled field strengths consistent with established safety limits.

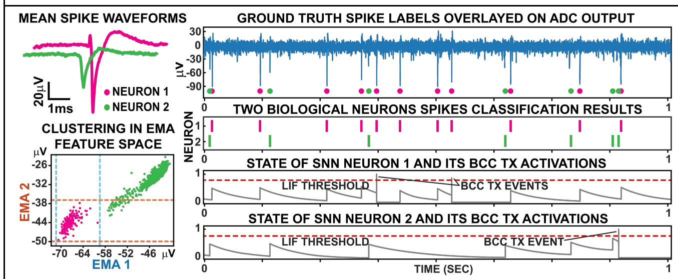


TABLE II. Comparison with spike classification and inference ICs.

	THIS WORK		ISSCC24 [1]	ISSCC25 [3]	ISSCC23 [4]	VLSI25 [2]
TECHNOLOGY (nm)	65		65	40	22 FDSOI	28
FEATURE EXTRACTION	EXPONENTIAL MOVING AVERAGE		TEMPORAL	SPATIAL	PEAK FSDE	STFE
SPIKE DETECTION AND CLASSIFICATION	LEARNABLE THRESHOLDS, Δs & BOOSTED DECISION STUMPS		THRESHOLD + BINNING + GAUSSIAN KERNEL	VALID COUNTER & NEO	NEO + GEO OSORT CLUSTERING	PEAK-DETECT + TEMPLATE MATCHING
SPIKE CLASSIF. ACCURACY	**93.1%** [a]		-	91.7% [b]	89.5% [a]	91.5% [a]
INFERENCE LOCATION	**SPLIT (LOCAL + REMOTE)**		LOCAL	N/A		
POST-SPIKE INFERENCE	**SNN & DNN WITH LOCAL 1ST LAYER**		DNC + LDA			
APPLICATION DATASET	MOTOR [14]	SPEECH [15]	HANDWRITING			
POST-SPIKE INFER. ACCURACY	90.4%	96.1%	90.8 - 91.3%			
TX DATA REDUCTION	1,066x	16,732x	968x [c] (7.68Mx)	2,688x [d]	39,272x [e]	-
DBE POWER/CHANNEL (µW)	**0.053**		0.44	0.074	1.78	2.37

a: QUIROGA DATASET b: MEA-REC DATASET c: CALCULATED UP TO DISTINCT NEURAL CODES FEATURES
d: SPIKE SORTING DATA REDUCTION ONLY e: DATA REDUCTION REPORTED ONLY FOR NEUROPIXEL DATASET
N/A: NOT APPLICABLE -: NOT REPORTED DBE: DIGITAL BACK-END **BOLD:** BEST PERFORMANCE

Fig. 9. Experimentally measured results from human cortical tissue (top) and comparison with SOA spike-processing ICs (bottom).

TABLE III. Comparison with state-of-the-art neural interface ICs.

		THIS WORK		ISSCC22 [10]	ISSCC24 [6]	ISSCC25 [11]	ISSCC23 [12]
SYSTEM	TECHNOLOGY (nm)	65		110/180	180	22	180
	SUPPLY (V)	1 / 2.5		2.3	1 / 1.25	1	1
	ON-CHIP PROCESSING	SPIKE CLASSIFICATION + 1ST LAYER OF SNN/ DNN		NO			
	WIRELESS POWERING	MAGNETOELECTRIC (ME)		BODY-COUPLED	INDUCTIVE	N/A	N/A
	STIMULATOR	YES		NO	YES	NO	NO
	SIGNAL TYPE	**RAW DATA**	**SPIKES**	SPIKES	LFP	LFP	LFP
	NO. INPUTS	1@16BITS	20@10BITS	4	1	1	1
	TOTAL POWER (µW)	51.2	121.5 [a]	644	270	4.62	9.3
ANALOG FRONT END (AFE)	IRN (µVrms)	**0.9**		6.6	7	1.32	4.69
	LNA NEF / PEF	**1.5 / 5.6**		- / -	- / -	3.47 / 12.05	- / -
	ADC ARCHITECTURE	NS SAR		CT-ΔΣ	CT-ΔΣ	NS SAR	NS SAR
	BW (kHz)	8		10	10	1	1
	POWER/INPUT (µW)	29	5.8	8.6	9.8	4.62	9.3
	SNDR (dB)	97	62	82.3	81.9	91.81	**99.1**
	ENOB (bits)	15.8	10	13.4	13	14.96	**16.2**
	FOM (dB) Schreier [b]	**181** [e]	153	173	172.2	175.2	179.4
	DYNAMIC RANGE (dB)	92	57	83.3	83.7	92.85	**101.6**
	VALIDATION	HUMAN/RODENT (EX/IN VIVO)		RODENT (IN VIVO)		HUMAN EEG	EEG
BCC DATA TX	DATA LINK TYPE	EVENT-DRIVEN BCC		BCC	RF	N/A	
	FREQUENCY (MHz) [c]	120		40.96	144		
	MAX DATA RATE (Mbps) [d] / @ MAX TX POWER (µW)	10 / 855		20.48 / 644	- / -		
	ENERGY/BIT (pJ) / DIST. (mm)	85.5 / 100		32 / 100	- / -		
	TX PWR/IN (µW) RAW DATA	21.9	20.5	-			
	TX PWR/IN (µW) WITH INFER. (DNN)	N/A	**0.00123**	N/A			

a: TOTAL POWER FOR 20 CH. WITH INFERENCE b: FOM (DB) SCHREIER = SNDR + 10 LOG(BW/PWR)
c: FREQ. DETERMINED BY OUTPUT EDGE SPACING d: TX OUTPUT GENERATED VIA DATA-EDGE COMBINING
e: 184 DB FOR ADC W/O LNA N/A: NOT APPLICABLE - : NOT REPORTED **BOLD**: BEST PERFORMANCE

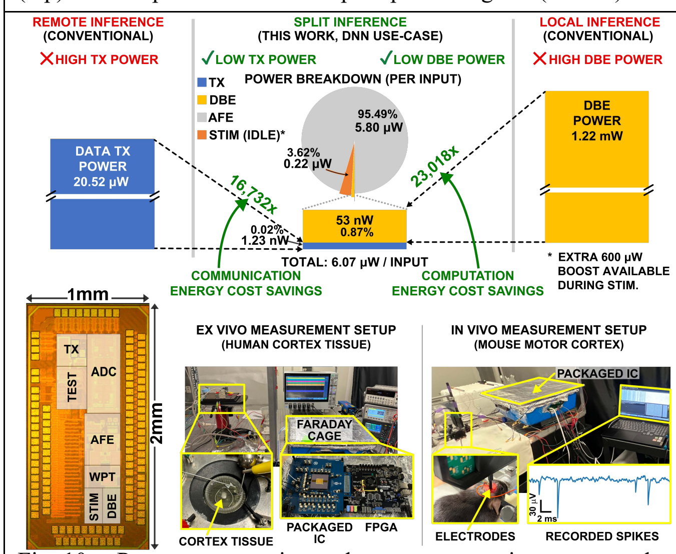


Fig. 10. Power consumption and energy cost savings compared to fully remote and fully local inference (top), and chip micrograph with experimental setups (bottom). Reported power includes inference logic, digital control, TX control, and serialization.

References

[5] J. Tang et al., ISSCC, 2021, pp. 470–472, doi: 10.1109/ISSCC42613.2021.9365859.
[6] L. Zhao et al., ISSCC, 2024, pp. 558–560, doi: 10.1109/ISSCC49657.2024.10454382.
[7] D. K. Piech et al., Nat BME, vol. 4, no. 2, pp. 207–222, 2020, doi: 10.1038/s41551-020-0518-9.
[8] W. Wang et al., JSSC, 2024, vol. 59, no. 11, pp. 3599–3611, 2024, doi: 10.1109/JSSC.2024.3464533.
[9] W. Wang et al., ISSCC, 2025, pp. 574–576, doi: 10.1109/ISSCC49661.2025.10904583.
[10] C. Lee et al., ISSCC, 2022, pp. 1–3, doi: 10.1109/ISSCC42614.2022.9731733.
[11] K. Jeong et al., ISSCC, 2025, pp. 276–278, doi: 10.1109/ISSCC49661.2025.10904745.
[12] G. Kim et al., ISSCC, 2023, pp. 484–486, doi: 10.1109/ISSCC42615.2023.10067844.
[13] M. C. W. V. Rossum, Neural Comp., vol. 13, no. 4, pp. 751–763, 2001, doi: 10.1162/089976601300014321.
[14] T. Brochier et al., Sci Data, vol. 5, no. 1, pp. 180055, 2018, doi: 10.1038/sdata.2018.55.
[15] S. K. Wandelt et al., Nat Hum Behav, vol. 8, no. 6, pp. 1136–1149, doi: 10.1038/s41562-024-01867-y.